

- 1 -

TITLE

"Co-translational protein folding simulations"

Background

5 A protein is a biomolecule consisting of one or more chains of amino acids, linked together by peptide bonds. Amino acids are organic compounds consisting of a central carbon atom, referred to as the α carbon, bonded to a carboxyl group, an amino group, a hydrogen atom, and a unique side chain R group. The unique R group determines the properties and characteristics of each amino acid.

10 Protein synthesis is a biological process occurring inside cells resulting in the production of new proteins. In general terms, it can be divided into transcription, the process of conversion of a gene into the template molecule messenger RNA, where information is conveyed by the sequence of nucleotides, and translation, the process of creation of the polypeptide chain by "translating" each codon (triplet of nucleotides) into the corresponding amino acid and through the formation of the peptide bond between amino acids, catalysed by the ribosome.

15 Protein folding is the process by which a protein acquires its native three-dimensional structure.

Understanding protein folding is one of the most challenging problems in molecular biology. The interest is not only in obtaining the final fold (generally referred to as structure prediction) but also in understanding the folding mechanism.

25 The available experimental techniques that probe proteins at different stages during the folding process (A.I. Bartlett and S. E. Radford, "An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms," *Nature Structural & Molecular Biology*, 16: 582–588, Jun. 2009, doi: <https://doi.org/10.1038/nsmb.1592>) and recent advances in the study of co-translational folding (A.A. Komar,

30

“Unraveling co-translational protein folding: Concepts and methods,”
Methods, 137: 71–81, Mar. 2018, doi:
<https://doi.org/10.1016/j.ymeth.2017.11.007>) have shed light on the
folding kinetics and thermodynamics involved in the actual folding
process. The folding process of many proteins takes the amino acid coil
through different intermediate states before stabilizing on the final folded
state. However, due to experimental limitations, detailed protein folding
pathways remain unknown.

A detailed model of the protein folding pathway remains a strongly felt
need.

Drawings

Figure 1: Comparison of experimental and simulation data. **A)**
Schematic representation of the force profile analysis sensor used for *in*
vitro translation experiments of the reference study (M. Liutkute, M. Maiti,
E. Samatova, J. Enderlein, and M. V. Rodnina, “Gradual compaction of
the nascent peptide during cotranslational folding on the ribosome,”
eLife, vol. 9, Oct. 2020, doi: <https://doi.org/10.7554/elife.60895>) part
of the figure is adapted from Liutkute et al., cited. It consists of the N-
terminal domain of the HemK protein (helices 1 to 5, H1-H5), the C-
terminal linker, the SecM arrest peptide and CspA. **B)** In black:
experimental values of the fraction of full-length construct product formed
during *in vitro* translation varying the number of residues of the construct
of wild-type HemK. In grey: predicted force profile (experimental values)
computed with data from cotranslational folding simulations of the folding
of the N-terminal domain of the HemK protein. Specifically, it represents
the increments in the fraction of native contacts varying the number of
synthesized residues ($\Delta\omega$). **C)** Schematic diagram of HemK N-terminal
domain compaction events during translation; grayscale tones as in
panel A, the figure is adapted from Liutkute et al., cited. **D)** Schematic
overlay of the HemK H1-H3 crystal structure (darker color) (PDB ID:

- 3 -

1T43) and ADR1 Zn finger domain crystal structure (lighter color) inside the peptide exit tunnel, 60–80 Å from the PTC, the figure is adapted from Liutkute et al., cited. **E)** Reconstruction of the compaction events occurring in the simulations according to the instant invention. Images were taken from frames belonging to one of the selected 37 trajectories.

Figure 2: Comparison of experimental and simulation data. In gray: ADR1a zinc finger domain crystal structure (PDB ID 2ADR) obtained from single-particle cryo-EM of co-translational folded ADR1a domain inside the *E. coli* ribosome exit tunnel. In light gray: portion of the 70S subunit that is employed in the simulations. In black: representation of a frame extracted from one of the simulated folding trajectories. The zinc finger was positioned inside the ribosome structure after proper alignment of the cryo-EM map to said ribosome structure. This allows to confirm that the first two helices of the HemK NTD dock inside the ribosome tunnel, at a height that is comparable to the one where the folded zinc finger structure is located.

Figure 3: Three sequential snapshots of structures of a given protein indicative of a folding trajectory, representing the starting configuration, an intermediate configuration, the final configuration. **A)** the folding trajectory is computed for an isolated protein. **B)** the folding trajectory is evaluated in the context of a molecular tunnel.

Figure 4: Representation of various configurations extracted from a folding trajectory of a given model protein chosen at regular time intervals along the trajectory. The x axis represents the “simulation time”, therefore the configuration on the far left represents the starting configuration of the trajectory, while the last state represents the native configuration of the given model protein. The y axis of the graph represents a generic simulation axis that has the same orientation as the longitudinal axis of what in some implementations of the method is referred to as a “molecular tunnel”, whereby the bottom of the y axis is towards the

“entrance” of such molecular tunnel while the top is oriented in the same direction as the “exit” of such tunnel. The portion of the protein that is above the dotted line coincides with the “relevant portion” of the protein.

5 **Figure 5:** Flow diagram illustrating an exemplary method of generating a folding trajectory in accordance with an aspect of the present invention.

Figure 6: Flow diagram illustrating an exemplary method of generating a folding trajectory in accordance with an aspect of the present invention.

10 **Figure 7:** Flow diagram illustrating an exemplary method of assigning the computed position in accordance with an aspect of the present invention.

Figure 8: Flow diagram illustrating an exemplary method of assigning the computed position in accordance with an aspect of the present invention.

15 **Figure 9: A), B)** Flow diagrams illustrating an exemplary method for identifying a target configuration in accordance with an aspect of the present invention.

Figure 10: Flow diagram illustrating an exemplary method to select a ligand in accordance with an aspect of the present invention.

Definitions

20 In this context, “protein synthesis” refers to the addition of amino acids, one after another, in the order in which their translating codons appear in the messenger RNA, to construct the polypeptide chain of a protein.

A “configuration” is the three-dimensional arrangement of a set of atoms.

25 A “folding trajectory” is an array of sequential configurations of the given protein, thus associated to the same sequence of amino acids.

A “distance between configurations” quantifies the spatial distance between two three-dimensional arrangements.

30 “Native configuration” means the three-dimensional arrangement of a protein, in its properly folded and/or assembled form, which is operative

- 5 -

and functional. It is one of the input data of the method according to the present invention.

“Given protein” identifies the protein to which the method is applied to, thus it consists of a specific sequence of amino acids which, in the method according to the present invention, is associated to one single specific native configuration.

“Starting configuration” refers to the three-dimensional arrangement of the given protein that is the starting point of a folding trajectory.

“Relevant portion” of the protein defines a portion of the protein that is relevant to the folding process at a given time and for the purposes of the method according to the present invention. The “relevant portion” of the protein is the part of said protein subjected to the folding biasing potential. This portion can be fully defined by two numbers representing the start and the end indices of a contiguous subsequence of the sequence of amino acids of the given protein. As an example, when a protein is analysed according to the instant method during its synthesis, said relevant portion is represented at the start by the portion of the protein that is already synthesized, at the end by the fully synthesized protein. In this embodiment, the relevant portion corresponds, in each iteration step, to the protein itself.

As an alternative, only a portion of said protein is analysed. This should be done when an already synthesized protein is analysed, or the protein is analysed during its synthesis. In this case, the relevant portion fictionally grows by addition of atoms that are already there on the protein and are added to the analysis during the process.

“Target configuration” indicates, among the configurations along the folding trajectory, the configuration suitable for binding drug-like molecules thus obtaining a modulation of the protein biological function.

A “protein contact map” is a two-dimensional matrix that represents a three-dimensional protein configuration. Each element is a function of the

- 6 -

distance between two residues or atoms of the protein, computed for all the possible amino acid residue pairs of the protein or for all the possible pairs of atoms within a subset of atoms of the protein (subset that could coincide with the protein itself).

5 **Description**

It forms a **first object** of the present invention a **method for determining the folding trajectory of a given protein coupled with its synthesis**.

10 The given protein comprises a sequence of amino acids and the folding trajectory comprises a sequence of configurations of said given protein, wherein a configuration is a set of atoms and their three-dimensional arrangement.

15 The folding trajectory produced by the method is such that the folding process is coupled and concurrent with the synthesis process of the polypeptide chain, in that the relevant portion of the protein comprises a number of atoms that is lower at the start of the trajectory compared to the end of the trajectory, increasing monotonically during the folding trajectory.

20 The method comprises a plurality of iterations, each iteration producing a new configuration starting from the previous one. In each iteration, atoms are optionally added to the configuration and atom positions are updated.

25 It forms a **second object** of the present invention a **method for identifying**, among said sequence of configurations, **a target configuration** suitable for being tested as target for drug discovery procedures.

30 It forms a **third object** of the present invention a **method performed by one or more data processing apparatus that computes a folding trajectory for a given protein**.

It forms a **fourth object** of the present invention a **method to select**

- 7 -

a ligand binding to a target protein, wherein said binding is observed between the ligand and a protein in one of the configurations obtained according to the invention.

5 In said first object, according to an embodiment, said method for determining the folding trajectory of a given protein comprises (**Figure 5**):

- a) generating a starting configuration;
- b) defining a metric to evaluate the distance between configurations;
- c) generating an array of protein configurations constituting the trajectory, comprising, at each of a plurality of iterations (**Figure 6**):
 - 10 – obtaining the protein configuration from the previous iteration;
 - optionally, adding one or more atoms to the relevant portion of the protein;
 - updating the position of one or more atoms of the protein configuration, in such a way that the distance between the newly
 - 15 generated configuration and the given final configuration gradually decreases; the gradual decrement is such that the next iteration could have a higher distance, but along the final trajectory for each frame it is possible to define an interval such that all the frames after that interval have a lower distance than
 - 20 that of the considered frame;
 - including at least some of said configurations into said folding trajectory;
 - halting after a pre-defined number of iterations and/or when the final configuration is reached and/or when the distance between
 - 25 the last configuration and the final configuration is below a given threshold.

In the context of the present invention, “optionally” adding one or more atoms to the relevant portion of the protein means that at each considered iteration the “adding one or more atoms to the relevant portion

30 of the protein” can be performed or not. In other words, within the context

- 8 -

of the present invention “adding one or more atoms to the relevant portion of the protein” is performed in at least one iteration, and in as many iterations as the number of atoms or residues to be added.

Therefore, in said first object, said method for determining the folding
5 trajectory of a given protein comprises:

- a) generating a starting configuration;
 - b) defining a metric to evaluate the distance between configurations;
 - c) generating an array of protein configurations constituting the trajectory, comprising, at each of a plurality of iterations:
 - 10 – obtaining the protein configuration from the previous iteration;
 - updating the position of one or more atoms of the protein configuration, in such a way that the distance between the newly generated configuration and the given final configuration gradually decreases; the gradual decrement is such that the next
15 iteration could have a higher distance, but along the final trajectory for each frame it is possible to define an interval such that all the frames after that interval have a lower distance than that of the considered frame;
 - including at least some of said configurations into said folding
20 trajectory;
 - halting after a pre-defined number of iterations and/or when the final configuration is reached and/or when the distance between the last configuration and the final configuration is below a given threshold;
- 25 wherein said step c) of generating an array of protein configurations constituting the trajectory further comprises, in at least one of said plurality of iterations, adding one or more atoms to the relevant portion of the protein, after said obtaining the protein configuration from the previous iteration and before said updating the position of one or more
30 atoms of the protein configuration; and wherein said relevant portion of

- 9 -

the protein comprises a number of atoms that is lower at the start of the trajectory compared to the end of the trajectory, increasing monotonically during the folding trajectory.

Said iterations are schematically represented in **Figure 3A**, wherein
5 three representative configurations, starting configuration, intermediate configuration, and final configuration, are illustrated for an isolated protein.

Said starting protein configuration is the configuration of the protein at the beginning of the trajectory, that is the first step of the abovementioned
10 plurality of iterations.

According to an embodiment of the method, the starting configuration is generated by assigning computed positions to one or more atoms of the given protein, wherein the computed positions are calculated from the positions of one or more atoms of a given template configuration, wherein
15 the protein of the given template configuration may have a different sequence from the given protein and, possibly, a different number of atoms and residues.

In an embodiment, said given template configuration is located inside a molecular tunnel, wherein a molecular tunnel is a given set of atoms surrounding a portion of a configuration of the given protein and this portion could be moved continuously outside the molecular tunnel without crossing the atoms of the tunnel.
20

In an embodiment, the molecular tunnel is the exit tunnel and the vestibule of the ribosome or the Sec61 translocation complex.

In an embodiment, said starting configuration is generated by assigning positions to all the atoms of the protein using a polypeptide template inside a molecular tunnel. In the next iterations, the residues of the protein will move out of the tunnel to continue the folding process. The relevant portion of the protein will grow, intended in terms of the
25 number of residues it is composed of, as a consequence of this shifting
30

of residues towards the exit of the molecular tunnel.

In an embodiment, said portion of the protein that is relevant to the folding is the portion of the protein that is exiting the vestibule of the ribosome, while in another embodiment, it is the portion of the protein that has exited the translocation channel and is folding in the Endoplasmic Reticulum lumen.

A representation of distinct configurations of a given protein along its folding trajectory is in **Figure 4**, wherein the bottom of the axis that is perpendicular to the “trajectory time” axis should be considered the entrance of said molecular tunnel and the top of said axis is oriented in the same direction as the exit of such tunnel.

In an embodiment, represented in **Figure 3B**, configurations of one or more additional chemical species are considered. As an example, said additional chemical species are selected from water molecules and/or sodium ions and/or chlorine ions.

In an embodiment, represented in **Figure 3B**, said folding trajectory further comprises a sequence of configurations of the molecular tunnel and/or of the additional chemical species, not only of the protein.

According to an embodiment of the method, the update of the positions of one or more atoms of the given protein and/or the molecular tunnel and/or the additional chemical species are carried out by defining a force field between the atoms and using an integration scheme for the equations of motion. The computation is done using a molecular dynamics engine such as NAMD, GROMACS, or OpenMM.

In an embodiment, represented in **Figure 7**, in each iteration the update of the positions comprises:

- obtaining the positions and velocities of the atoms that are the result of the computation at the previous iterative step;
- computing the gradient, with respect to the positions of the atoms, of a predefined force field function that depends on the positions

of the atoms;

- computing the gradient with respect to the positions of the atoms of a biasing potential that acts on the relevant portion of the protein and promotes the formation of native contacts and/or disfavours the formation of non-native contacts;
- updating the positions and the velocities of the atoms using the previously computed gradients with respect to the positions of the atoms.

According to an embodiment, the biasing potential used to promote native contacts and/or disfavour non-native contacts is defined using the contact map of the final conformation and/or the contact map of the current conformation.

According to an embodiment, said biasing potential is based on the distance between the current contact map and the native contact map of the protein.

In an embodiment, said biasing potential further comprises a function defined on the positions of a subset of the atoms that has a minimum in a particular configuration of the atoms, wherein the configuration of the atoms defining the minimum can change between updating iterations; and/or a function defined on one or more absolute coordinates of the coordinate system.

According to an embodiment of the method, a second biasing potential can be added to help the protein flow through the molecular tunnel during the simulation. The parameterization of this potential defines the speed and the modality of the expulsion out of the tunnel.

In an embodiment, said biasing potential term is used to obtain configurations increasingly outside of the molecular tunnel.

In an embodiment, one or more of the biasing potential terms are active only in a subset of the updating iterations.

In an embodiment, the updating of the positions of the atoms is done

by a molecular dynamic engine such as NAMD, GROMACS, OpenMM.

In an embodiment, the metric used to define the distance between configurations is defined:

- using a contact map;
- 5 – and/or using the Root Mean Squared Value (RMSD);
- and/or using the fraction of native contacts (Q).

According to an embodiment of the method, the molecular tunnel is fed using an algorithm to add new residues to the polypeptide chain during the simulation. In the example implementation the residues are added with the dihedrals in extended conformation one at a time.

In an embodiment, adding one or more atoms to the protein configuration comprises adding the atoms of an amino acid at the C-terminus of the given protein configuration, comprising (**Figure 8**):

- obtaining the most C-terminal amino acid of the current configuration of the given protein;
- 15 – obtaining the list of amino acids of the given protein;
- finding, in the list of amino acids of the given protein, the next amino acid with respect to the most C-terminal amino acid of the current configuration of the given protein, wherein next is intended from the N-terminus to the C-terminus order;
- 20 – obtaining the set of atoms of the next amino acid;
- assigning computed positions to the set of atoms of the next amino acid;
- adding the set of atoms of the next amino acid and their positions to the current configuration of the given protein.
- 25

In an embodiment, said computed positions are assigned by a method comprising (**Figure 8**):

- a) Obtaining the three-dimensional coordinates of the backbone of a dipeptide composed of two residues in extended conformation. In an implementation example said positions are retrieved from pre-
- 30

- 13 -

- computed values;
- b) Obtaining the three-dimensional coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein;
- 5 c) Rotating and translating the three-dimensional coordinates of the backbone of the dipeptide so that the coordinates of the backbone of the first residues of the dipeptide are aligned with the coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein. In an implementation
- 10 example said alignment is the one minimizing the Root Mean Square Distance between the aligned atoms;
- d) Obtaining the three-dimensional coordinates of the next amino acid. In an implementation example, said coordinates are retrieved from a pre-computed library of residues in extended
- 15 conformation. In another implementation example said coordinates are retrieved from the corresponding residue of the native configuration of the given protein;
- e) Rotating and translating the three-dimensional coordinates of the next amino acid so that the coordinates of the backbone of the
- 20 next amino acids are aligned with the coordinates of the backbone of the second residue of the dipeptide. In an implementation example said alignment is the one minimizing the Root Mean Square Distance between the aligned atoms;
- f) Rotating the three-dimensional coordinates of the next amino acid
- 25 around the axis connecting the N atom and the C-alpha carbon (i.e., phi dihedral) until no clashes with atoms of the current configuration of the given protein and/or atoms of the molecular tunnel and/or atoms of additional molecular species are detected. In an implementation example a clash is defined when the
- 30 distance between a pair of atom coordinates is below a threshold;

- 14 -

- g) If no rotation avoiding all clashes can be found, performing an energy minimization. In an implementation example the energy minimization is performed with a molecular dynamic engine such as NAMD, GROMACS, OpenMM;
- 5 h) In an embodiment when water molecules and/or ions are present, removing water molecules and/or ions close to the next residue. In an implementation example the water molecules and/or ions to remove are selected based on a distance threshold with the atoms of the next residue;
- 10 i) In an embodiment when ions are present and the net electrical charge of all the species is not zero, adding or removing ions to restore neutrality.

According to an embodiment, the biasing potential to promote native contacts is applied to the system with a ratchet and pawl technique. In an implementation example, the applied bias potential is the following:

15

$$U_{\text{ratchet}}(t) = \max \left(r_{\text{fold}}(X(t)) - \min_{\tau < t} (r_{\text{fold}}(X(\tau))), 0 \right)^2$$

where r_{fold} , C_{ij} and σ are defined as follows:

20

$$r_{\text{fold}} = k_{\text{fold}} \sum_{ij} \left(C_{ij}(X) - C_{ij}(X_{\text{native}}) \right)^2$$

$$C_{ij}(X) = a_{ij}^0 \sigma \left(\frac{|X_i - X_j|}{r_{ij}^0} \right)$$

25

$$\sigma(\xi) = \frac{1 - \xi^6}{1 - \xi^{10}}$$

where X_i are the tridimensional coordinates of the i -th atom of the protein, k_{fold} , a_{ij}^0 , r_{ij}^0 , are parameters determining the strength and the profile of the biasing forces.

According to an embodiment, the biasing potential used to drive the polypeptide chain along the molecular tunnel is a harmonic force on a

30

- 15 -

selection of atoms of the chain where the equilibrium position of the harmonic potential of each atom changes during the simulation to pull the chain in the desired direction. In an implementation example, the potential is applied on alpha carbons, and it is like the following:

$$U_i^{\text{pull}}(X, t) = \frac{1}{2} k_i(t) \left(X_i(t) - X_i^{\text{ref}}(t) \right)^2$$

where X_i are the three-dimensional coordinates of the i -th atom of the selection of atoms and X_i^{ref} are the three-dimensional coordinates of the equilibrium positions for the i -th atom.

According to an embodiment, the equilibrium position of the harmonic potential for the i -th atom in the chain is determined with the following formula:

$$X_i^{\text{ref}}(t) = X_{i - \left\lfloor \frac{t}{\tau} \right\rfloor + 1}^{\text{template}}(t = 0)$$

where X_i^{template} are three-dimensional coordinates of atoms on a given template polypeptide, I is a parameter to determine the equilibrium position of the atoms at $t=0$ and τ is the simulation time that passes between two shifts in the equilibrium positions of the harmonic potential.

In this implementation example, the bias potential used to drive the polypeptide is applied only on atoms for which $i - \left\lfloor \frac{t}{\tau} \right\rfloor$ is greater or equal to 0.

According to an embodiment of the method, the elastic constant of the harmonic potential for the i -th atom in the chain is determined with the following formula:

$$k_i(t) = k_0 \theta \left(i - \frac{t}{\tau} \right)$$

where θ is a step function that is 0 when the argument is negative and 1 otherwise.

It is here further claimed, in a second object, a method for identifying a target configuration, said method comprising (**Figures 9A, 9B**):

- a) obtaining one or more folding trajectories;
- b) clustering the configurations of the one or more folding trajectories based on features computed on the atomic positions of the

- 16 -

configurations considering only the atoms belonging to the relevant portion of each configuration. In an implementation example, these features are the contact maps.

- 5 c) selecting one or more clusters that are relevant folding intermediates, wherein a cluster is considered a relevant folding intermediate if it contains configurations that belong to a high percentage of folding trajectories and/or contain a relevant percentage of configurations out of the total of available configurations from the one or more collected folding trajectories;
- 10 d) computing the intermediate lifetime for the one or more relevant clusters, where the lifetime is computed considering the synthesis rate of the protein in the species of interest; in an implementation example for human proteins the synthesis rate is set to 4 amino acids per second and the predicted intermediate lifetime should be no less than 1 second;
- 15 e) selecting a configuration out of each of the clusters selected with the previous steps, wherein said selected configuration:
- comprises amino acids of said relevant portion (criterion A);
 - has at least a pocket that is not present in the native state and is only accessible in the intermediate state by a putative small molecule (criterion B), wherein said pocket:
 - has good physico-chemical characteristics (criterion C), as computed by some pocket scouting software.
 - is stable (criterion D), i.e., if the configuration is taken as the
- 20
- 25 pocket physico-chemical parameters are consistent throughout the simulation.

According to an embodiment of the method, selecting the target configuration comprises performing all the steps of the previous embodiment and additionally (**Figure 9B**):

30

- 17 -

- starting a molecular dynamics simulation from the selected configuration, that is at least 500ns long and where the backbone of the protein is restrained with harmonic forces where the minimum of the potential for each atomic coordinate coincides with the starting position of such atom;
- selecting, from the sequence of configurations obtained through the molecular dynamic's simulations, a configuration that satisfies the following characteristics:
 - comprises amino acids of said relevant portion (criterion A);
 - the selected configuration has at least a pocket that is not present in the native state and is only accessible in the intermediate state by a putative small molecule (criterion B);
 - said pocket has good physico-chemical characteristics (criterion C), as computed by some pocket scouting software;
 - said pocket is stable (criterion D), i.e., if the configuration is taken as the starting configuration of a molecular dynamic simulation, the pocket physico-chemical parameters are consistent throughout the simulation.

Regarding the physico-chemical characteristics, different embodiments of the method provides that such scoring parameters comprise "SiteScore", and/or "Dscore", and/or "DrugScore" and or pocket balance.

These parameters are per se known, in the field of modelling and characterization of protein and protein intermediates, and are based on a mix of values, related to different properties, suitable to evaluate the "druggability" of a protein native state or a protein folding intermediate.

In fact, such scoring parameters derive from known evaluation software packages.

For example, SiteMap (T. A. Halgren, 2009 "Identifying and Characterizing Binding Sites and Assessing Druggability," J. Chem. I

Information and Modeling, 49:377–389, doi:
<https://doi.org/10.1021/ci800324m>) predicts a site score (SiteScore) and
druggability score (DScore) through a linear combination of only three
single descriptors: the size of the binding pocket, its enclosure, and a
5 penalty for its hydrophilicity.

Another example is DoGSiteScorer (A. Volkamer, D. Kuhn, T.
Grombacher, F. Rippmann, and M. Rarey 2012 “Combining Global and
Local Measures for Structure-Based Druggability Predictions,” J. Chem.
Information and Modeling, 52:360–372, doi:
10 <https://doi.org/10.1021/ci200454v>) which also generates a druggability
score (DrugScore) which range from zero to one.

In an implementation example, said pocket scouting software is
DogSiteScorer and the pocket physico-chemical characteristics are
considered good when the descriptor volume is at least 350 Å³,
15 hydrophobicity greater than 0.6 and smaller than 0.9, depth of at least 13,
drug score and simple score at least 0.5.

A particularly preferred embodiment of the second object of the
present invention is a method for identifying a target configuration
suitable for being tested as target for drug discovery procedures, wherein
said step a) of obtaining one or more folding trajectories of a given protein
20 is performed via the method disclosed above as the first object of the
present invention, i.e. a method comprising:

- A. generating a starting configuration;
- B. defining a metric to evaluate the distance between configurations;
- 25 C. generating an array of protein configurations constituting the
trajectory, comprising, at each of a plurality of iterations:
 - obtaining the protein configuration from the previous iteration;
 - updating the position of one or more atoms of the protein
configuration, in such a way that the distance between the newly
30 generated configuration and the given final configuration

gradually decreases; the gradual decrement is such that the next iteration could have a higher distance, but along the final trajectory for each frame it is possible to define an interval such that all the frames after that interval have a lower distance than that of the considered frame;

- 5 – including at least some of said configurations into said folding trajectory;
- halting after a pre-defined number of iterations and/or when the final configuration is reached and/or when the distance between
- 10 the last configuration and the final configuration is below a given threshold;

wherein said step C) of generating an array of protein configurations constituting the trajectory further comprises, after said obtaining the protein configuration from the previous iteration and before said updating

15 the position of one or more atoms of the protein configuration, in at least one of said plurality of iterations, adding one or more atoms to the relevant portion of the protein; and wherein said relevant portion of the protein comprises a number of atoms that is lower at the start of the trajectory compared to the end of the trajectory.

20 Therefore, in this preferred embodiment, said method for identifying a target configuration suitable for being tested as target for drug discovery procedures then may further comprise the following steps, performed after said step a):

25 b) clustering the configurations of the one or more folding trajectories based on features computed on the atomic positions of the configurations considering only the atoms belonging to the relevant portion of each configuration. In an implementation example, these features are the contact maps.

30 c) selecting one or more clusters that are relevant folding intermediates, wherein a cluster is considered a relevant folding

- 20 -

- intermediate if it contains configurations that belong to a high percentage of folding trajectories and/or contain a relevant percentage of configurations out of the total of available configurations from the one or more collected folding trajectories;
- 5 d) computing the intermediate lifetime for the one or more relevant clusters, where the lifetime is computed considering the synthesis rate of the protein in the species of interest; in an implementation example for human proteins the synthesis rate is set to 4 amino acids per second and the predicted intermediate lifetime should
- 10 be no less than 1 second;
- e) selecting a configuration out of each of the clusters selected with the previous steps, wherein said selected configuration:
- comprises amino acids of said relevant portion (criterion A);
 - has at least a pocket that is not present in the native state
 - 15 and is only accessible in the intermediate state by a putative small molecule (criterion B), wherein said pocket:
 - has good physico-chemical characteristics (criterion C), as computed by some pocket scouting software.
 - is stable (criterion D), i.e., if the configuration is taken as the
 - 20 starting configuration of a molecular dynamic simulation, the pocket physico-chemical parameters are consistent throughout the simulation.
- Preferably, said adding one or more atoms to the relevant portion of the protein comprises adding the atoms of an amino acid at the C-
- 25 terminus of the given protein configuration, comprising:
- obtaining the most C-terminal amino acid of the current configuration of the given protein;
 - obtaining the list of amino acids of the given protein;
 - finding, in the list of amino acids of the given protein, the next
 - 30 amino acid with respect to the most C-terminal amino acid of the

- 21 -

current configuration of the given protein, wherein next is intended from the N-terminus to the C-terminus order;

- obtaining the set of atoms of the next amino acid;
- assigning computed positions to the set of atoms of the next amino acid;
- adding the set of atoms of the next amino acid and their positions to the current configuration of the given protein.

Even more preferably, said adding one or more atoms to the relevant portion of the protein is performed via a method comprising:

- i. obtaining the three-dimensional coordinates of the backbone of a dipeptide composed of two residues in extended conformation. In an implementation example said positions are retrieved from pre-computed values;
- ii. obtaining the three-dimensional coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein;
- iii. rotating and translating the three-dimensional coordinates of the backbone of the dipeptide so that the coordinates of the backbone of the first residues of the dipeptide are aligned with the coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein. In an implementation example said alignment is the one minimizing the Root Mean Square Distance between the aligned atoms;
- iv. obtaining the three-dimensional coordinates of the next amino acid. In an implementation example, said coordinates are retrieved from a pre-computed library of residues in extended conformation. In another implementation example said coordinates are retrieved from the corresponding residue of the native configuration of the given protein;
- v. rotating and translating the three-dimensional coordinates of the next

- 22 -

- amino acid so that the coordinates of the backbone of the next amino acids are aligned with the coordinates of the backbone of the second residue of the dipeptide. In an implementation example said alignment is the one minimizing the Root Mean Square Distance between the aligned atoms;
- 5
- vi. rotating the three-dimensional coordinates of the next amino acid around the axis connecting the N atom and the C-alpha carbon (i.e., phi dihedral) until no clashes with atoms of the current configuration of the given protein and/or atoms of the molecular tunnel and/or
- 10
- atoms of additional molecular species are detected. In an implementation example a clash is defined when the distance between a pair of atom coordinates is below a threshold;
- vii. if no rotation avoiding all clashes can be found, performing an energy minimization. In an implementation example the energy minimization
- 15
- is performed with a molecular dynamic engine such as NAMD, GROMACS, OpenMM;
- viii. optionally, in an embodiment when water molecules and/or ions are present, removing water molecules and/or ions close to the next residue. In an implementation example the water molecules and/or
- 20
- ions to remove are selected based on a distance threshold with the atoms of the next residue;
- ix. optionally, in an embodiment when ions are present and the net electrical charge of all the species is not zero, adding or removing ions to restore neutrality.
- 25
- It forms a **fourth object** of the present invention a **method to select a ligand (Figure 10)**, wherein a ligand is a compound capable to bind in a selective manner to a target protein, as a not limiting example it is a drug, the method comprising:
- a) Obtaining the three-dimensional configuration of a target protein,
- 30
- wherein a target protein comprises a sequence of amino acids and

- 23 -

- a given final configuration;
- b) computing one or more folding trajectories of said target protein, according to the above-described method;
- c) selecting one or more configurations of the folding trajectory;
- 5 d) evaluating an interaction of one or more candidate ligands with the selected configurations of the folding trajectory;
- e) selecting one or more of the candidate ligands as the ligand, wherein said selection is based on the results of said evaluation.

10 In an embodiment, said interaction of one of the candidate ligands comprises determining an interaction score for the candidate ligand, wherein the interaction score comprises a measure of an interaction between the candidate ligand and the target protein in the selected configuration.

15 In a further embodiment, it is here described a system comprising one or more computers and one or more storage devices storing instructions that when executed by the one or more computers cause the one or more computers to perform the operations of the method according to the present invention.

20 In a further embodiment, it is here described one or more computer storage media storing instructions that when executed by one or more computers cause the one or more computers to perform the operations of the respective method according to the present invention.

Advantages

25 The method according to the present invention allows to effectively add residues to the protein during the folding process, mimicking the peptide bond formation operated at the peptidyl transferase center (PTC). Such an implementation saves simulation time, avoiding the performance penalty due to the larger number of protein and solvent atoms to simulate in the method according to the state of the art.

30 Moreover, the method allows to monitor the folding process as it happens

during protein synthesis, therefore better replicating the biological context.

The restraints guiding the synthesized residues can have a different harmonic constant depending on their height in the exit tunnel, to facilitate the formation of tertiary contacts inside the ribosome. Moreover, they can extend above the structure present at the simulation's beginning, allowing such a structure to start at the bottom of the tunnel without losing the ability to guide the newly added residues toward the exit.

Experimental section

Example 1: method validation

The results published in Liutkute et al., cited, employing an indirect method to describe the dynamics of co-translational folding of the N-terminal domain of HemK (HemK methyltransferase), have been used for comparative purpose to validate the method according to the present invention.

Liutkute et al. obtained a "force profile" detailing the pulling force exerted by the nascent chain during its synthesis and this information was used to provide a coarse-grained interpretation of how different tertiary structures fold inside the ribosome. In this study, the force-profile sensor represented in **Figure 1A** has been used.

To compare the results obtained by the here claimed in silico method with the in vitro available results, 145 co-translational simulations of the HemK protein have been performed. The synthesis of new residues was stopped after the first 120 N-terminal residues, while only residues GLU2 to GLN101 were subject to the biasing potential. Residues were synthesized every 128000 rMD integration steps and after stalling, the simulation was extended for additional 1.632×10^6 integration steps, resulting in a total of 13.792×10^6 integrations steps for each trajectory.

Of these 145 trajectories, 81 fold correctly based on the RMSD filtering criterion of 3.75 Å applied to the N-terminal domain (residues 2 to 101).

An additional filtering criterion was applied to identify the trajectories which start folding inside the exit tunnel since these are the relevant trajectories that allow for a comparison with the experimental results.

5 In particular, only those trajectories in which contact between helix 1 and helix 2 is formed before exiting the ribosome were retained. A total of 37 trajectories (corresponding to 45.7% of the folding ones) met this criterion and were used for the analysis.

10 The fraction of native contacts Q for each frame of these trajectories was computed and the values of the frames where the number of synthesized residues is the same were averaged. The average number of native contacts per residue ω was then obtained by dividing these average Q values by the number of newly synthesized residues (total number of residues minus number of residues at the start of the simulation). Then, $\Delta\omega$ values were computed as differences between ω values that are 5 residues apart (e.g., $\Delta\omega(\#res) = \omega(\#res + 5) - \omega(\#res)$).
15 $\Delta\omega$ can be interpreted as a proxy of the pulling force exerted by the folding chain, as it is reasonable to suppose that folding events significantly incrementing w would exert a higher pulling force on the chain. $\Delta\omega$ can therefore be compared with the values of the force profile analysis coming from the in vitro translation experiments of Liutkute et al., cited. The results of the comparison are shown in **Figure 1B**. As a side note, to effectively compare the two, the $\Delta\omega$ values have been shifted to the right by 15 residues. This means that our first value is aligned in the graph to the experimental force profile value at 15 residues.
20 The reason for this offset is in part given by the 5 residues gap in the computation of $\Delta\omega$, and in part by the fact that our simulation starts with a nascent chain that is already 24 residues long. If we used the information of the length of the SecM sequence (17 residues) and assumed that our synthesis point corresponds to the actual position of the PTC, we could argue that the total shift of our data with respect to the
25
30

experimental ones would be $5 + 24 - 17 = 13$. Indeed, since the location of our synthesis point may not be perfectly superimposable on the PTC, we computed the optimal shift value by calculating the Pearson correlation coefficient between the experimental values and our data, shifted by different amounts; the maximum correlation is achieved with a shifting of 15. The error in the estimate of the $\Delta\omega$ values was obtained with a bootstrapping procedure. In particular, the average of the Q values was computed after resampling 1000 times the Q vector. The resampling was performed on the trajectories, selecting the Q values belonging to 37 trajectories chosen randomly with replacement from the original set of 37 trajectories. From these 1000 Q vectors, $\Delta\omega$ vectors were computed. **Figure 1B** displays their mean, and the filled curve is their standard deviation.

The two curves clearly exhibit the same behaviour. They define the same zones of higher and lower exerted force, zones likely corresponding to different events characterizing the folding process. The peaks of the graph, corresponding to four major structural changes, are highlighted by a shaded background as in Liutkute et al. Small discrepancies between the two curves are, in any case, to be expected, as neither of them is a direct measurement of the underlying observable they aim to represent (i.e., the pulling force).

A model for the folding of the HemK NTD was then derived by Liutkute et al., assigning compaction events to the four peaks of the force profile analysis. This model is shown in **Figure 1C**. For comparison, we reconstructed the sequence of compaction events occurring in our simulations, extracting frames at the specific times in the trajectory where the number of residues (after applying the shift) matches the residues on the x-axis of graph of **Figure 1B**. All the frames belong to one specific trajectory of the selected 37. Of the 11 depicted states, only 2 do not perfectly match the ones predicted by the authors. In the 6th

conformation, corresponding to a protein length of 51 residues, the third helix (residues 36-41) is not docked to helices 1 and 2. It is worth mentioning, however, that the first 3 helices can correctly fold before exiting the vestibule, as predicted by the authors. On the other hand, in the 9th conformation (protein length: 79 residues), the non-native compact state does not form. One possible explanation for this observation could be that the ratchet dynamics disfavours the formation of non-native contacts.

The possibility for small domains to fold inside the ribosome exit tunnel has been confirmed by Nilsson et al. ("Cotranslational Protein Folding inside the Ribosome Exit Tunnel," *Cell Reports*, 12:1533–1540, 2015, doi: <https://doi.org/10.1016/j.celrep.2015.07.065>) by employing different experimental techniques. In particular, the authors obtained a cryo-electron microscopy density map of a small zinc-finger domain protein (ADR1a, PDB: 2ADR) folded inside the exit tunnel of the E. coli ribosome. This experiment suggests the ability of the ribosome to accommodate the folding of the first 3 helices of HemK, as postulated by the authors of Liutkute et al. To confirm this possibility, the authors superimposed the first 3 helices bundle with the zinc finger structure (**Figure 1D**). In our simulations, we indeed observe the folding of the 3 helices in the region occupied by the zinc finger (**Figure 1E**). **Figure 2** represents one of these simulations superimposed to the zinc finger, after proper alignment of the ribosome to the cryo-EM map.

CLAIMS

1. A method for determining the folding trajectory of a given protein comprising:
- a) generating a starting configuration;
 - 5 b) defining a metric to evaluate the distance between configurations;
 - c) generating an array of protein configurations, comprising, at each of a plurality of iterations:
 - obtaining the protein configuration from the previous iteration;
 - optionally, adding one or more atoms to the relevant portion of the protein;
 - 10 – updating the position of one or more atoms of the protein configuration, in such a way that the distance between the newly generated configuration and the given final configuration gradually decreases; the gradual decrement is such that the next iteration could have a higher distance, but along the final trajectory for each frame it is possible to define an interval such that all the frames after that interval have a lower distance than that of the considered frame;
 - halting after a pre-defined number of iterations and/or when the final configuration is reached and/or when the distance between the last configuration and the final configuration is below a given threshold;
 - 20 – including at least some of said configurations into said folding trajectory;
- 25 wherein the relevant portion of the protein comprises a number of atoms that is lower at the start of the trajectory compared to the end of the trajectory.
2. The method according to claim 1, wherein said relevant portion coincides at any given time along the trajectory with the protein itself,
- 30 wherein the protein itself consists of all the amino acids of the given

protein that are present in each configuration of the trajectory and comprises a number of atoms that is lower at the start of the trajectory compared to the end of the trajectory.

- 5 3. The method according to claim 1, wherein said starting configuration is generated by assigning computed positions to one or more atoms of the given protein, wherein the computed positions are calculated from the positions of one or more atoms of a given template configuration, wherein the given template protein consists of a configuration of a protein that may have a different sequence from
10 the given protein and possibly a different number of atoms and residues.
4. The method according to claim 3, wherein said given template protein is located inside a molecular tunnel, preferably said molecular tunnel being the exit tunnel and the vestibule of the ribosome or the Sec61
15 translocation complex.
5. The method according to any one of the claims 1-4, wherein said starting configuration is generated by assigning positions to all the atoms of the protein using a polypeptide template inside a molecular tunnel.
- 20 6. The method according to any one of the claims 1-5, wherein configurations of one or more additional chemical species are considered.
7. The method according to any one of the claims 1-6, wherein the iterations of the atoms of the given protein and/or the molecular
25 tunnel and/or the additional chemical species are carried out by defining a force field between the atoms and using an integration scheme for the equations of motion.
8. The method according to any one of the claims 1-7, wherein said iterations comprise:
30 – obtaining the positions and velocities of the atoms that are the

- 30 -

- result of the computation at the previous iterative step;
- defining a force field function that depends on the positions of the atoms;
 - determining the gradient of the force field function with respect to the positions of the atoms;
 - updating the positions and the velocities of the atoms using the gradient with respect to the positions of the atoms.
- 5
9. The method according to claim 8, wherein said force field defined on the relevant portion comprises an additional biasing potential to promote the formation of native contacts and/or disfavour the formation of non- native contacts.
- 10
10. The method according to claim 1, wherein said step of adding one or more atoms to the relevant portion of the protein comprises adding the atoms of an amino acid at the C-terminus of the given protein configuration, comprising:
- 15
- obtaining the most C-terminal amino acid of the current configuration of the given protein;
 - obtaining the list of amino acids of the given protein;
 - finding, in the list of amino acids of the given protein, the next amino acid with respect to the most C-terminal amino acid of the current configuration of the given protein, wherein next is intended from the N-terminus to the C-terminus order;
 - obtaining the set of atoms of the next amino acid;
 - assigning computed positions to the set of atoms of the next amino acid;
 - adding the set of atoms of the next amino acid and their positions to the current configuration of the given protein.
- 20
- 25
11. The method according to claim 1, wherein said step of adding one or more atoms to the relevant portion of the protein is performed with a method comprising:
- 30

- 31 -

- a) obtaining the three-dimensional coordinates of the backbone of a dipeptide composed of two residues in extended conformation;
 - b) obtaining the three-dimensional coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein;
 - 5 c) rotating and translating the three-dimensional coordinates of the backbone of the dipeptide so that the coordinates of the backbone of the first residues of the dipeptide are aligned with the coordinates of the backbone of the most C-terminal amino acid of the current configuration of the given protein;
 - 10 d) obtaining the three-dimensional coordinates of the next amino acid;
 - e) rotating and translating the three-dimensional coordinates of the next amino acid so that the coordinates of the backbone of the next amino acids are aligned with the coordinates of the backbone of the second residue of the dipeptide;
 - 15 f) rotating the three-dimensional coordinates of the next amino acid around the axis connecting the N atom and the C-alpha carbon (i.e., phi dihedral) until no clashes with atoms of the current configuration of the given protein and/or atoms of the molecular tunnel and/or atoms of additional molecular species are detected;
 - 20 g) if no rotation avoiding all clashes can be found, performing an energy minimization;
 - h) optionally, when water molecules and/or ions are present, removing water molecules and/or ions close to the next residue;
 - 25 i) optionally, when ions are present and the net electrical charge of all the species is not zero, adding or removing ions to restore neutrality.
- 12.A method for identifying a target configuration suitable for being tested as target for drug discovery procedures, said method
- 30

- 32 -

comprising:

- a) obtaining one or more folding trajectories, according to one of the claims 1-11;
- 5 b) clustering the configurations of the one or more folding trajectories based on features based on the atomic positions of the configurations computed on the relevant portion of each configuration;
- 10 c) selecting one or more clusters that are relevant folding intermediates, wherein a cluster is considered a relevant intermediate if it contains configurations that belong to a high percentage of folding trajectories and/or contain a relevant percentage of configurations out of the total of available configurations from the one or more collected folding trajectories;
- 15 d) computing the intermediate lifetime for the one or more relevant clusters, where the lifetime is computed considering that the synthesis rate in humans is 4 aa per second and that the intermediate lifetime should be no less than 1 second;
- 20 e) selecting a configuration out of each of the clusters selected with the previous steps, wherein said selected configuration:
 - has at least a pocket that is not present in the native state and is only accessible in the intermediate state by a putative small molecule, wherein said pocket:
 - comprises amino acids of said relevant portion;
 - is stable, i.e., if the configuration is taken as the starting
 - 25 configuration of a molecular dynamic simulation, the pocket physico-chemical parameters are consistent throughout the simulation.
- 30 **13.** The method according to claim 12, wherein said step e) further comprises evaluating physico-chemical characteristics of said pocket, by using a pocket scouting software and said pocket physico-

chemical characteristics are considered good when the descriptor volume is at least 350 Å³, hydrophobicity greater than 0.6 and smaller than 0.9, depth of at least 13, drug score and simple score at least 0.5.

- 5 **14.** A method to select a ligand, wherein a ligand is a compound capable to bind in a selective manner to a target protein, said method comprising:
- 10 a) obtaining the three-dimensional configuration of a target protein, wherein a target protein comprises a sequence of amino acids and a given final configuration;
- b) computing the folding trajectory of said target protein, according to one of the claims 1-11;
- c) selecting one or more configurations of the folding trajectory, according to claim 12;
- 15 d) evaluating an interaction of one or more candidate ligands with the selected configurations of the folding trajectory;
- e) selecting one or more of the candidate ligands as the ligand, wherein said selection is based on the results of said evaluation.
- 20 **15.** A system comprising one or more computers and one or more storage devices storing instructions that when executed by the one or more computers cause the one or more computers to perform the operations of the method according to one of the claims 1-11.